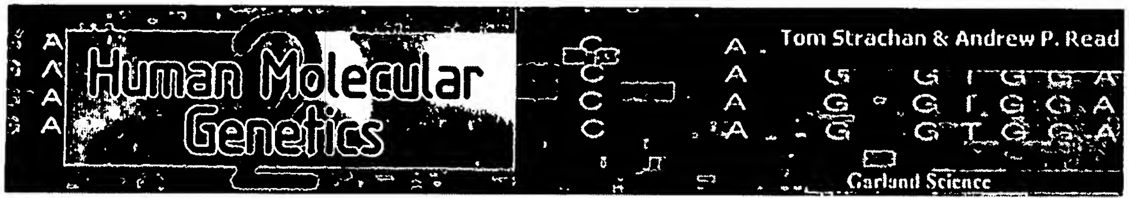
Exhibit 1



5.1. Preparation of nucleic acid probes

In standard nucleic acid hybridization assays the probe is labeled in some way. Nucleic acid probes may be made as single-stranded or double-stranded molecules (see *Figure 5.1*), but the working probe must be in the form of single strands.

Conventional DNA probes are isolated by cell-based DNA cloning or by PCR. In the former case, the starting DNA may range in size from 0.1 kb to hundreds of kilobases in length and is usually (but not always) originally double-stranded. PCR-derived DNA probes have often been less than 10 kb long and are usually, but not always, originally double-stranded. Conventional DNA probes are usually labeled by incorporating labeled dNTPs during an *in vitro* DNA synthesis reaction (see Section 5.1.1).

RNA probes can conveniently be generated from DNA which has been cloned in a specialized plasmid vector (Melton *et al.*, 1984). Such vectors normally contain a phage promoter sequence immediately adjacent to the multiple cloning site. An RNA synthesis reaction is employed using the relevant phage RNA polymerase and the four rNTPs, at least one of which is labeled. Specific labeled RNA transcripts can then be generated from the cloned insert (see Section 5.1.1).

Oligonucleotide probes are short (typically 15–50 nucleotides) single-stranded pieces of DNA made by chemical synthesis: mononucleotides are added, one at a time, to a starting mononucleotide, conventionally the 3′ end nucleotide, which is bound to a solid support. Generally, oligonucleotide probes are designed with a specific sequence chosen in response to prior information about the target DNA. Sometimes, however, oligonucleotide probes are used which are **degenerate** in sequence. Typically this involves parallel syntheses of a set of oligonucleotides which are identical at certain nucleotide positions but different at others. Oligonucleotide probes are often labeled by incorporating a ^{32}P atom or other labeled group at the 5′ end (see next section).

5.1.1. DNA and RNA can conveniently be labeled in vitro by incorporation of nucleotides (or nucleotide components) containing a labeled atom or chemical group

Although, in principle, DNA and RNA can be labeled *in vivo*, by supplying labeled deoxynucleotides to tissue culture cells, this procedure is of limited general use; it has been restricted largely to preparing labeled viral DNA

Search

GO

● This book ○ All books

○ PubMed

from virus-infected cells, and studying RNA processing events. A much more versatile method involves *in vitro* labeling: the purified DNA, RNA or oligonucleotide is labeled *in vitro* by using a suitable enzyme to incorporate labeled nucleotides. Two major types of procedure have been widely used:

- **Labeling of new strands during*in vitro*DNA or RNA synthesis.** In this type of procedure, DNA or RNA polymerase is used to make labeled DNA or RNA copies of a starting DNA. The *in vitro* DNA or RNA synthesis reaction requires that at least one of the four nucleotide precursors carries a labeled group. Labeling of DNA by *in vitro* DNA synthesis is normally accomplished using one of three methods: nick-translation (this section); random primed labeling (this section); or PCR-mediated labeling (see Section 6.1.1). Labeling of RNA generally is carried out using an *in vitro* transcription system (see further on in this section).
- **End-labeling.** This type of procedure involves addition of a labeled group to one or a few terminal nucleotides. It is less widely used, but is useful for a number of procedures, including labeling of single-stranded oligonucleotides (see below) and restriction mapping. Inevitably, because only one or a very few labeled groups are incorporated, the specific activity (the amount of radioactivity incorporated divided by the total mass) of the labeled DNA is much less than that for probes in which there has been incorporation of several labeled nucleotides along the length of the DNA.

Labeling DNA by nick translation

The nick-translation procedure involves introducing single-strand breaks (*nicks*) in the DNA, leaving exposed $3'$ hydroxyl termini and $5'$ phosphate termini. The nicking can be achieved by adding a suitable endonuclease such as pancreatic deoxyribonuclease I (DNase I). The exposed nick can then serve as a start point for introducing new nucleotides at the $3'$ hydroxyl side of the nick using the DNA polymerase activity of *E. coli* DNA polymerase I at the same time as existing nucleotides are removed from the other side of the nick by the $5' \rightarrow 3'$ exonuclease activity of the same enzyme. As a result, the nick will be moved progressively along the DNA ('translated') in the $5' \rightarrow 3'$ direction (see *Figure 5.2A*). If the reaction is carried out at a relatively low temperature (about 15° C), the reaction proceeds no further than one complete renewal of the existing nucleotide sequence. Although there is no net DNA synthesis at these temperatures, the synthesis reaction allows the incorporation of labeled nucleotides in place of the previously existing unlabeled ones. ↑ TOP

Random primed DNA labeling

The random primed DNA labeling method (sometimes known as oligolabeling) (Feinberg and Vogelstein, 1983) is based on hybridization of a mixture of all possible hexanucleotides: the starting DNA is denatured and then cooled slowly so that the individual hexanucleotides can bind to suitably complementary sequences within the DNA strands. Synthesis of new complementary DNA strands is primed by the bound hexanucleotides and is catalyzed by the Klenow subunit of DNA polymerase I (which contains the polymerase activity in the absence of associated exonuclease activities). DNA synthesis occurs in the presence of the four dNTPs, at least one of which has a labeled group (see *Figure 5.2B*). This method produces labeled DNAs of high specific activity. Because all sequence combinations are represented in the hexanucleotide mixture, binding of primer to template DNA occurs in a random manner, and labeling is uniform across the length of the DNA. ↑ TOP

End-labeling of DNA

Single-stranded oligonucleotides are usually end-labeled using polynucleotide kinase (kinase end-labeling). Typically, the label is provided in the form of a ^{32}P at the γ-phosphate position of ATP and the polynucleotide kinase catalyses an exchange reaction with the 5′-terminal phosphates (see *Figure 5.3A*). The same procedure can also be used for labeling double-stranded DNA. In this case, fragments carrying label at one end only can then be generated by cleavage at an internal restriction site, generating two differently sized fragments which can be separated by gel electrophoresis and purified.

Larger DNA fragments can be end-labeled by various alternative methods. **Fill-in end-labeling** (*Figure 5.3B*) is one popular approach, and uses the *Klenow subunit* of *E. coli* DNA polymerase. Again, fragments carrying label at one end only can be generated by restriction cleavage and size fractionation. An alternative PCR-based method is primer-mediated 5′ end-labeling (see Section 6.1.1). ↑ TOP

Labeling of RNA

The preparation of labeled RNA probes (riboprobes) is most easily achieved by *in vitro* transcription of insert DNA cloned in a suitable plasmid vector. The vector is designed so that adjacent to the multiple cloning site is a phage promoter sequence, which can be recognized by the corresponding phage RNA polymerase. For example, the plasmid vector pSP64 contains the bacteriophage SP6 promoter sequence immediately adjacent to a multiple cloning site (see *Figure 5.4*). The SP6 RNA polymerase can then be used to initiate transcription from a specific start point in the SP6 promoter sequence, transcribing through any DNA sequence that has been inserted into the multiple cloning site. By using a mix of NTPs, at least one of which is labeled, high specific activity radiolabeled transcripts can be generated (*Figure 5.4*). Bacteriophage T3 and T7 promoter/RNA polymerase systems are also used commonly for generating riboprobes. Labeled sense and antisense riboprobes can be generated from any gene cloned in such vectors (the gene can be cloned in either of the two orientations) and are widely used in tissue *in situ* hybridization (Section 5.3.4). ↑ TOP

5.1.2. Nucleic acids can be labeled by isotopic and nonisotopic methods

Isotopic labeling and detection

Traditionally, labeling of nucleic acids has been conducted by incorporating nucleotides containing radioisotopes. Such radiolabeled probes contain nucleotides with a radioisotope (often ^{32}P, ^{33}P, ^{35}S or ^{3}H), which can be detected specifically in solution or, much more commonly, within a solid specimen (autoradiography - see *Box 5.1*).

The intensity of an autoradiographic signal is dependent on the intensity of the radiation emitted by the radioisotope, and the time of exposure, which may often be long (one or more days, or even weeks in some applications). ^{32}P has been used widely in Southern blot hybridization, dot-blot hybridization, colony and plaque hybridization (see below) because it emits high energy β-particles which afford a high degree of sensitivity of detection. It has the disadvantage, however, that it is relatively unstable (see *Table 5.1*). Additionally, its high energy β-particle emission can be a disadvantage under circumstances when fine physical resolution is required to interpret the resulting image unambiguously. For this reason, radionuclides which provide less energetic β-particle radiation have been preferred in certain procedures, for example ^{35}S-labeled and ^{33}P-labeled nucleotides for DNA sequencing and tissue *in situ* hybridization, and ^{3}H-labeled nucleotides for chromosome *in situ* hybridization. ^{35}S and ^{33}P have moderate half-lives while ^{3}H has a very long half-life. However, the latter isotope is disadvantaged by its comparatively low energy β-particle emission which necessitates very long exposure times.

^{32}P-labeled and ^{33}P-labeled nucleotides used in DNA strand synthesis labeling reactions have the radioisotope at the α-phosphate position, because the β- and γ-phosphates from dNTP precursors are not incorporated into the growing DNA chain. Kinase-mediated end-labeling, however, uses [γ-^{32}P]ATP (see *Figure 5.3A*). In the case of ^{35}S-labeled nucleotides which are incorporated during the synthesis of DNA or RNA strands, the NTP or dNTP carries a ^{35}S isotope in place of the O^{-} of the α-phosphate group. ^{3}H-labeled nucleotides carry the radioisotope at several positions. Specific detection of molecules carrying a radioisotope is most often performed by autoradiography (see *Box 5.1*). TOP

Nonisotopic labeling and detection

Nonisotopic labeling systems involve the use of nonradioactive probes. Although developed only comparatively recently, they are becoming increasingly popular and are finding increasing applications in a variety of different areas (Kricka, 1992). Two types of non-radioactive labeling are conducted:

- **Direct nonisotopic labeling**, where a nucleotide which contains the label that will be detected is incorporated. Often such systems involve incorporation of modified nucleotides containing a fluorophore (*Figure 5.5A*), a chemical group which can fluoresce when exposed to light of a certain wavelength (fluorescence labeling - see *Box 5.2*).
- **Indirect nonisotopic labeling**, usually featuring the chemical coupling of a modified reporter molecule to a nucleotide precursor. After incorporation into DNA, the reporter groups can be specifically bound by an affinity molecule, a protein or other ligand which has a very high affinity for the reporter group. Conjugated to the latter is a marker molecule or group which can be detected in a suitable assay (*Figure 5.6*). The reporter molecules on modified nucleotides need to protrude sufficiently far from the nucleic acid backbone to facilitate their detection by the affinity molecule and so a long carbon atom *spacer* is required to separate the nucleotide from the reporter group.

Two indirect nonisotopic labeling systems are widely used:

- The biotin-streptavidin system utilizes the extremely high affinity of two ligands: biotin (a naturally occurring vitamin) which acts as the reporter, and the bacterial protein streptavidin, which is the affinity molecule. Biotin and streptavidin bind together extemely tightly with an affinity constant of 10^{-14}, one of the strongest known in biology. Biotinylated probes can be made easily by including a suitable biotinylated nucleotide in the labeling reaction (see *Figure 5.7*).

- **Digoxigenin** is a plant steroid (obtained from *Digi-talis* plants) to which a specific antibody has been raised. The digoxigenin-specific antibody permits detection of nucleic acid molecules which have incorporated nucleotides containing the digoxigenin reporter group (see *Figure 5.7*).

A variety of different marker groups or molecules can be conjugated to affinity molecules such as streptavidin or the digoxigenin-specific antibody. They include various fluorophores (see *Box 5.2*), or enzymes such as alkaline phosphatase and peroxidase which can permit detection via colorimetric assays or chemical luminescence assays, etc. ↑ TOP

Exhibit 2

Published online July 8, 2004

Nucleic Acids Research, 2004, Vol. 32, No. 12 ***e99***
doi:10.1093/nar/gnh099

Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression

Cheng-Chung Chou, Chun-Houh Chen[1], Te-Tsui Lee and Konan Peck*

Institute of Biomedical Sciences and [1]Institute of Statistical Science, Academia Sinica, Taipei, Taiwan 115, Republic of China

Received December 5, 2003; Revised March 5, 2004; Accepted June 21, 2004

ABSTRACT

Gene-specific oligonucleotide probes are currently used in microarrays to avoid cross-hybridization of highly similar sequences. We developed an approach to determine the optimal number and length of gene-specific probes for accurate transcriptional profiling studies. The study surveyed probe lengths from 25 to 1000 nt. Long probes yield better signal intensity than short probes. The signal intensity of short probes can be improved by addition of spacers or using higher probe concentration for spotting. We also found that accurate gene expression measurement can be achieved with multiple probes per gene and fewer probes are needed if longer probes rather than shorter probes are used. Based on theoretical considerations that were confirmed experimentally, our results showed that 150mer is the optimal probe length for expression measurement. Gene-specific probes can be identified using a computational approach for 150mer probes and they can be treated like long cDNA probes in terms of the hybridization reaction for high sensitivity detection. Our experimental data also show that probes which do not generate good signal intensity give erroneous expression ratio measurement results. To use microarray probes without experimental validation, gene-specific probes ~150mer in length are necessary. However, shorter oligonucleotide probes also work well in gene expression analysis if the probes are validated by experimental selection or if multiple probes per gene are used for expression measurement.

INTRODUCTION

DNA microarrays are widely regarded as a powerful tool for large-scale gene expression measurements. The two main DNA microarray platforms are cDNA and oligonucleotide microarrays. cDNA microarrays are made with long double-stranded DNA molecules generated by enzymatic reactions such as PCR (1), while oligonucleotide microarrays employ oligonucleotide probes spotted by either robotic deposition or *in situ* synthesis on a solid substrate (2). It should be noted that, in this article, the immobilized DNA molecules are referred to as the probes and the labeled gene transcripts for hybridization as the targets, as suggested in Vol. 21, Supplement, Chipping Forecast, *Nature Genetics* 1999.

If the probes are not optimized for sequence specificity, both types of DNA microarrays can generate false-positive data due to non-specific cross-hybridization to highly similar sequences, gene families (3,4), or alternatively spliced variants (5). Cross-hybridization of one probe to several targets occurs more often with cDNA microarrays than with gene-specific oligonucleotide microarrays. In this article, 25–30mer probes are short oligonucleotide probes and 50–80mer probes are long oligonucleotide probes. Long DNA probes refer to probes of 100–150mer in length. cDNA probes are derived from cDNA clones and are ⩾500 bases in length. Literature reports (3,4) have shown that, if the targets have >70–80% global sequence homology to the cDNA probe, they can hybridize indiscriminately to the cDNA probe. In addition, high local sequence similarity between different sequences also causes significant cross-hybridization (3). Long oligonucleotide probes are also prone to cross-hybridization. For instance, any non-target sequence showing 75% similarity to a 50mer oligonucleotide probe results in cross-hybridization (6), and the same is true for non-target sequences showing 70% similarity to a 60mer probe (7). These observations have suggested that the percentage sequence homology is a reasonable predictor of cross-hybridization (4). To overcome this cross-hybridization problem, a general practice adopted by several laboratories is to design oligonucleotide probes targeting regions of low sequence similarity (6–8).

However, the use of oligonucleotide probes to replace cDNA probes in microarrays for expression profiling has generated discussion about the discordant results obtained using these two types of probes (9,10), the optimal oligonucleotide probe length and the number of oligonucleotide probes needed to obtain reliable expression data for a gene (11). Literature data (7,11) indicate that longer oligonucleotides (e.g. 60–80mers) provide significantly better detection sensitivity than shorter probes (e.g. 25 or 30mers). However, these long oligonucleotide probe microarrays use only one probe per gene, despite the fact that oligonucleotide hybridization is highly sequence dependent (12). It has been reported that

*To whom correspondence should be addressed. Tel: +886 2 2652 3072; Fax: +886 2 2785 8594; Email: konan@ibms.sinica.edu.tw oligonucleotide probes binding to different regions of a gene yield different signal intensities (2,7,13), and it is difficult to predict whether an oligonucleotide probe will bind efficiently to its target sequence and yield a good hybridization signal on the basis of sequence information alone (14). Because of this, multiple probes per gene have been used in oligonucleotide array designs to obtain reliable quantitative information of gene expression (2,7,13). Early versions of *in situ* synthesized 20mer oligonucleotide arrays employed 20 probe pairs per gene to provide statistically reliable quantification (2). On the basis of accumulated experimental results, probes that do not yield good hybridization signals were excluded to reduce the number of probes per gene. Five probes per gene has been suggested to be a good number for 30mer probes (11). When the best single probe for a gene was selected experimentally from eight 60mer candidate probes, it successfully detected gene expression at low levels (7). An experimental study showed that >25% of the probes in a set of 15 357 80mer probes were incapable of producing a usable hybridization signal (http://www.clontech.com/archive/APR02UPD/BDAtlas.shtml). To ensure good hybridization signals, every oligonucleotide probe should be tested experimentally, but the large-scale screening process is extremely time-consuming and costly.

In view of the aforementioned facts, since the major difference between different microarray platforms is the immobilized probes, the differences in the expression measurement results obtained can be attributed to differences in the probes. The sequence and length of a probe are critical factors for DNA microarray performance. Different probe formats show different degrees of sequence-dependent hybridization variation, different hybridization sensitivities and specificities, and a different number of probes per gene. In this article, we present an approach that can be used to investigate these issues and its use in evaluating probe lengths and the number of probes per gene for optimal microarray analysis of gene expression.

MATERIALS AND METHODS

Probe design

The human UniGene database (build #153) was used to derive source sequences for probe design. The gene sequences were filtered to remove repetitive elements using the RepeatMasker program (A. F. A. Smit and P. Green, http://ftp.genome.washington.edu/RM/RepeatMasker.html), then aligned using the BLAST program (15) with source sequences compiled from the UniGene database and the TIGR gene indices. The probe selection strategy was mainly based on frequently used criteria (6–8) for setting a sequence similarity threshold for specificity and a range of GC content for uniform hybridization reaction. Any part of a gene having 75% local sequence similarity in a 50 base window with other genes was masked. The remaining unmasked sequences were considered unique sequences of the gene. Within these unique sequences 100 and 150mer probes were designed using the Primer3 program (Unix version 0.9, S. Rozen and H. J. Skaletsky, http://www-genome.wi.mit.edu/genome_software/other/primer3.html). The GC content was calculated for every oligonucleotide probe ranging from 25 to 70mer selected from within the unique sequences of the gene transcript using a one-base-shift tiling method. We selected the gene-specific probes with a GC content between 45 and 55% and used the Mfold program (16) to further screen for probes that bind to the target sequences with the maximal Gibbs free energy of secondary structure.

Array fabrication

The 25mer, 50mer and 70mer oligonucleotides and all PCR primer pairs were synthesized on a 1536-channel DNA synthesizer (17) constructed in-house. Unless otherwise specified, all the oligonucleotide probes were modified with spacer and 5′ amino-linker moieties. The spacer was positioned between the probe sequence and the linker to extend the probe sequence away from the surface for better access to the target molecules. The spacer was a single unit of hexa-ethyloxy-glycol and was added to the probe using DMT-hexa-ethyloxy-glycol CED phosphoramidite (ChemGenes) and standard phosphoramidite synthesis chemistry. The amino-linker was coupled to the spacer using the same phosphoramidite synthesis chemistry and was used to tether probes by their 5′ ends to the slide surface with covalent bonding. Probes 100mer or longer were generated by PCR using a sense primer with the same spacer and 5′ amino-linker modifications. The concentrations of the oligonucleotide probes and primers were measured on a UV spectrophotometer at 260 nm (SpectraMax Plus 384, Molecular Devices). All the unique cytochrome P450 gene probes were generated by PCR amplification of the 3′-untranslated regions (3′-UTRs) or exons in genomic DNA derived from human placenta tissue. The high-performance liquid chromatography (HPLC)-purified oligonucleotide probes and column-purified (QIAquick, Qiagen) long DNA probes were dissolved in 150 mM sodium phosphate buffer (pH 8.5). The probes were spotted in duplicate and covalently immobilized via the 5′ end to surface-activated slides (SurModics) using an in-house constructed arrayer. The subsequent processing of slides was performed according to the manufacturer's instructions. For long DNA probe arrays, the double-stranded probes generated by PCR were rendered single-stranded by heating in boiling water for 2 min. Quality assessment of array printing was performed using a fluorescent dye, SYTO 61 (Molecular Probes), to directly measure the amount of DNA retained on the slide surface (18).

Preparation of *in vitro*-transcribed polyadenylated RNA

Five plant genes (*rbcl*, *rca*, ga*4*, *hat4* and *atps*) and two human genes (*gapdh* and β-*actin*) were PCR-amplified with gene-specific primers then cloned into the pCITE-4a(+) vector (Novagen). Thirty-one cytochrome P450 (CYP) family gene members were amplified by PCR using gene-specific primers (PCR amplicon size 800–1000 bp), with a T7 promoter sequence attached to the forward primer and $(dT)_{25}$ attached to the reverse primer. *In vitro*-transcribed polyadenylated RNAs, generated using T7 RNA polymerase (Epicenter), were purified on an RNeasy column (Qiagen). All the *in vitro*-derived RNAs were quantified using a BioAnalyzer (Agilent).

Sample labeling, hybridization and image analysis

For direct cDNA labeling, 2 μg of polyadenylated RNA sample was labeled by reverse transcription with 300 μM Cy3- or Cy5-dUTP (Amersham) at 42°C for 2 h in a 30 μl reaction mixture containing 1 μg of oligo(dT) (18–20mer, Invitrogen), 1.5 μg of random hexamer (Invitrogen), 200 μM dTTP, 500 µM dCTP, dATP and dGTP (Amersham), 400 U Superscript II Reverse Transcriptase (Invitrogen), 1 µM DTT and 1× reverse transcription buffer. After completion of the labeling reaction, 15 µl of 0.1 M NaOH/2 mM EDTA solution was added to stop the reaction and degrade the RNA. The reaction mixture was neutralized by the addition of 15 µl of 0.1 M HCl. The labeled cDNA was purified using size exclusion chromatography (YM-30 column, Millipore).

Hybridization of fluorescently labeled cDNA to the slide was performed at 42°C for 16–18 h in 10 µl of 5× SSC, 0.1% SDS and 50% formamide under an 18 × 18 mm cover-slip in a sealed chamber. After hybridization, the arrays were sequentially washed in 2× SSC, 0.1% SDS at 42°C for 5 min, in 0.2× SSC at room temperature for 1 min, and in 0.1× SSC at room temperature for 1 min. After drying by centrifugation, the arrays were scanned with a GenePix 4000B scanner (Axon Instruments). Array image acquisition and signal analysis were performed using GenePix Pro 4.0 software.

Measurement bias

To estimate the number of probes needed per gene to obtain statistically significant expression results, we randomly picked different numbers of probes from the available probes (population probes) for every gene. Chauvenet's criterion was used for rejection of outlying data points (19). Probes with an intensity deviation greater than a critical value, defined as the deviation divided by the sample standard deviation, were regarded as statistical outliers and excluded. The remaining probes were used for the calculation of the mean intensity. We used the average intensity (sample mean of signals) to represent the hybridization signal for a gene. The same approach was used to calculate the population signal of the population probes for a gene. The average expression measurement bias between the sample and population signal intensities (measurement bias hereafter) is defined as

$$\frac{\sum_N \sum_M |(\text{sample signal} - \text{population signal})/\text{population signal}|}{N \times M},$$

where M and N are, respectively, the number of iterative samplings and the number of genes used for the calculation.

The loss function

In order to identify the optimal probe length based on three factors, cross-hybridization (CH), measurement bias (MB) and the coefficient of variation (CV) of the signal intensity (SI), a loss function (20–23) was defined as $L(\text{CH, MB, SI}) = w_1 \cdot \text{CH}^2 + w_2 \cdot \text{MB}^2 + (1 - w_1 - w_2) \cdot \text{SI}^2$, where $(0 \leq w_1, w_2 \leq 1)$ are the weights for CH, MB and SI. The weights, w_1 and w_2, for CH, MB and SI were set at 1/3 (equal weight) to balance these three components. The quadratic loss function was then simplified to $L(\text{CH, MB, SI}) = \text{CH}^2 + \text{MB}^2 + \text{SI}^2$. This is a deterministic function, since no probability is involved.

RESULTS

Optimization of probe length: theoretical and empirical considerations

Effect of probe length on hybridization signal intensity variation. We analyzed two sets of literature data on genome-wide expression profiling, one for *Escherichia coli* (15 PM and MM pairs of 25mer probes per gene transcript, with ~80% of the genes having positive signals for at least 8 of the probe pairs) (13) and the other for *Saccharomyces cerevisiae* (8 60mer probes per gene) (7). The coefficient of variation, defined as the standard deviation divided by the mean of the hybridization signals for all the probes of each gene, was used as a measure of intensity variation. Based on statistical principles of sampling, the smaller the CV for the hybridization signals, the fewer the probes per gene needed for reliable gene expression measurement. The CV for the genome-wide average hybridization intensity was smaller for the 60mer probe set (0.55) than for the 25mer probe set (1.06).

To further explore the CV for signal intensity (hybridization efficiency) for longer probe lengths, we carried out experimental studies in which the same amount of seven Cy3-labeled cDNAs (five plant genes and two human genes) derived by *in vitro*-transcription of full-length gene clones (see Materials and Methods) were individually hybridized to an array of 336 probes. For each of the seven genes, there were six different probe lengths, with eight probes for each length. Probe selection was based on the algorithm described in the Materials and Methods. Figure 1A shows that the average

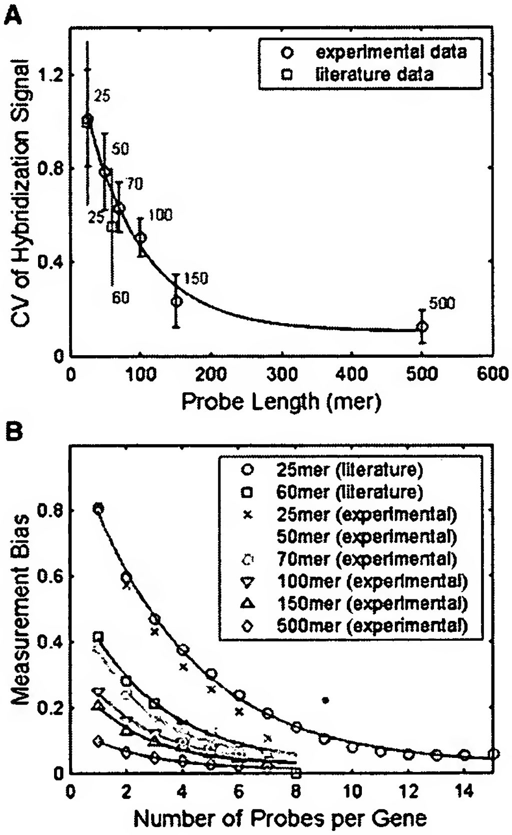


Figure 1. Effect of probe length and the number of probes per gene on expression measurement. (A) Effect of probe length on the variation (CV) in the hybridization signal using different length probes for the same genes. The experiments were repeated at least four times and the fluorescence signals were measured by Cy3 intensity. (B) Effect of the number of probes per gene on measurement bias.

CV for the hybridization intensity decreased monotonically with increasing probe length. These experimental results showed a reasonable agreement with the literature data of genome-wide expression measurements (7,13). The CV for hybridization intensity did not reach 0, but leveled off at a value of 0.1 for probes with 500 bases. This residual CV may result from spotting variation introduced by the arrayer, as the CV for the intensity of SYTO 61 staining of probes ranged from 7 to 12%.

Effect of the number of probes per gene on measurement bias. To evaluate measurement bias, we used the previous two literature datasets and the experimental data for the seven Cy3 labeled cDNAs. Figure 1B shows that the measurement bias (M = 100) decreased with an increase in the number of probes per gene. Fewer probes per gene were required for the longer probes to achieve the same bias as shorter probes. Although we used only seven genes in the experiments, the results corresponded quite well with the genome-wide literature data for *E.coli* and *S.cerevisiae* (7,13). These results show that a single 60–70mer probe per gene selected using a computational approach without experimental validation may not accurately measure the 'true' gene expression signal intensity and that multiple oligonucleotide probes per gene are required to obtain statistically reliable gene expression data.

Influence of probe length on cross-hybridization. Although long DNA or cDNA probes provide a low measurement bias and signal intensity variation, they often exhibit poor discrimination and hybridize to similar sequences, causing cross-hybridization problems. Cross-hybridization is also dependent on the stringency of the hybridization protocols. Several reports on the use of microarrays and commercial user manuals [(6,7,11,24), Amersham 30mer Uniset Bioarray, and MWG 50mer array protocol manuals] show that hybridization at 42°C in 30–50% formamide-based buffer works well for spotted DNA microarrays with probe lengths from 25mers to >1 kb (cDNA probes). Using fixed hybridization conditions in this study, the extent of cross-hybridization depends mainly on the sequence similarity between the sample targets and the microarray probes.

It is well known that the position of mismatched bases on short probes dramatically affects hybridization behavior (25). For instance, a single central mismatch in a 25mer probe completely eliminated the hybridization signal (13) and a single mismatch at 50 bases from the end of the probe attached to the slide surface of a 60mer microarray significantly reduced the hybridization signal (7). We therefore took these aforementioned mismatch positions on probes (25 and 60mer) into consideration in the cross-hybridization prediction below.

To compute the effect of probe length on cross-hybridization, we performed an *in silico* prediction of probe cross-hybridization to non-target human transcripts. We first randomly sampled 1000 genes from the UniGene database (build #153) and selected probes with lengths ranging from 25 to 1000 bases for each gene. Probe sequences of a particular length were chosen by a tiling method, one base at a time, along each gene. These probes were then aligned with the TIGR gene index (HGI version 9) using the BLAST program to analyze whether they would cross-hybridize to non-target sequences. A probe was considered prone to cross-hybridization if its sequence composition met the criteria for sequence similarity (see Materials and Methods) and mismatch positions described above.

By performing computations for over 10^6 probes for the 1000 genes, we found that among those prone to cross-hybridization probes, as predicted by the sequence similarity computation, only 3.2% of the 25mer probes and 3.5% of the 60mer probes had single mismatches at the center (25mer probes) or 10 bases from the 5′ end (60mer probes). These results show that sequence similarity is the dominant factor in genome-wide cross-hybridization computation. The percentage of the probes for the 1000 genes that would cross-hybridize versus probe length is shown in blue in Figure 2A. The plot shows that randomly chosen 50mer probes give the minimum cross-hybridization.

Probe length optimization. In addition to the cross-hybridization curve, Figure 2A combines the CV curve in Figure 1A with extrapolation to 1000 bases and the average measurement bias data extracted from the Figure 1B data for a single probe per gene. As shown in Figure 2A, hybridization signal variation and measurement bias work in opposition to cross-hybridization for selecting an optimal probe length. Thus, the

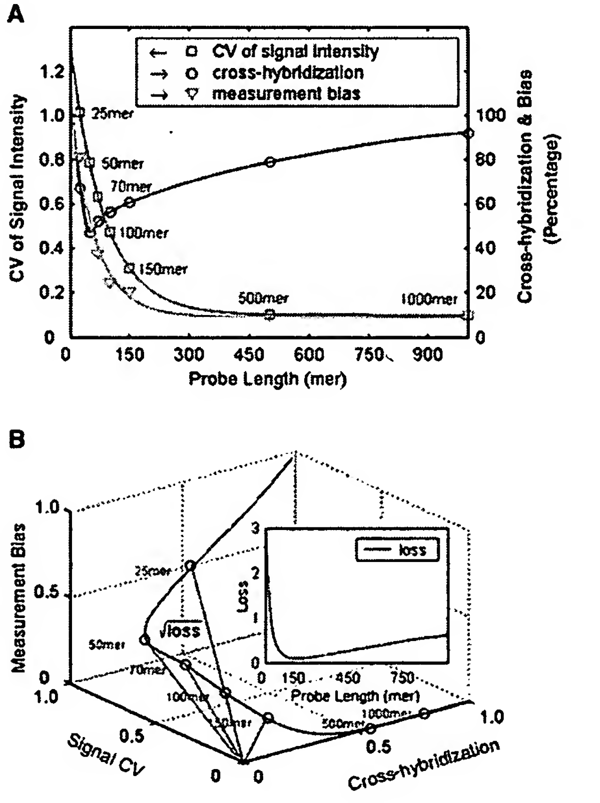


Figure 2. Determination of the optimal probe length. (A) Plots of cross-hybridization, measurement bias for one probe per gene, and the CV of the signal intensity versus probe length. All the solid lines were fitted by regression analysis of the data points. The scale for cross-hybridization and measurement bias is on the right of the figure. (B) The data from the three curves in (A) were plotted against each other in an XYZ plot, the minimum and maximum values for all parameters being normalized, respectively, to 0 and 1. The Euclidean distance between the curve and the origin represents the square root of the loss. The inset shows a plot of loss versus probe length to locate the minimum.

optimal probe length is a compromise length chosen to minimize these conflicting effects. We have developed a mathematical approach for the optimization. A statistical loss function (see Materials and Methods), defined as the square of the Euclidean distance to the origin (Figure 2B), was used to identify the optimal probe length, a smaller value of the loss function reflecting that the compound negative effects were closer to zero. The quadratic curve for loss function versus probe length up to 1000mer, displayed in the inset in Figure 2B, showed that the minimum loss of 0.085 was seen with a 150–160mer probe. The loss function calculation provides a reference for optimal probe length selection. This conclusion was supported by the experimental validation below.

Optimization of probe length: experimental evaluation

Effects of probe length and probe concentration on hybridization sensitivity. In vitro-transcribed polyadenylated RNA for a plant gene, *rbcl*, was directly labeled with Cy3 dye by reverse transcription, then hybridized to an array containing *rbcl* probes of five different lengths ranging from 25 to 150 bases at various probe concentrations (0.2–100 μM). The probe sequence used for the indicated length was the best probe experimentally selected from the eight probes for each length using the same procedure for obtaining the data shown in Figure 1A. A 500mer cDNA clone-derived probe at a concentration of 1 μM was spotted on the array for comparison. Figure 3A shows curves of the signal intensity versus probe concentration for different probe lengths. The experimental results showed that long DNA probes gave a more intense hybridization signal than the long oligonucleotide probes at every probe concentration tested, but that higher concentrations of long oligonucleotide probes gave the same signal intensity as lower concentrations of long DNA probes. For example, 50–70mer probes at a concentration of 20–40 μM gave a similar hybridization signal to 150 or 500mer probes at a concentration of 1 μM. These results show that long DNA probes, which extend farther away from the slide surface than oligonucleotide probes, are more accessible to free target molecules for hybridization. Although oligonucleotide probes are less accessible, a high surface density resulting from a high spotting concentration largely improves their poor hybridization signal intensity. These findings are in agreement with those in previous reports (11,12,26).

Spacer effect. Figure 3B shows that increasing the probe length by addition of spacers of longer length enhanced the hybridization intensity. The spotting concentrations of 25–70mer oligonucleotides and 100–150mer PCR-derived single-stranded probes were 10 and 1 μM, respectively, as suggested by the slide manufacturer. The spacer effect was greater for 50–70mer probes, but negligible for long DNA probes (100–150mer); for the 25mer shown, much longer spacers were required to show a significant effect (data not shown). Previous reports have indicated that the addition of a spacer has a large effect on the hybridization signal intensity for 15–30mer oligonucleotides, but that the signal decreases with spacer length after an optimal length is reached (26–28). The hybridization intensities for the 100 and 150mer probes eventually reached the same plateau level (Figure 3B).

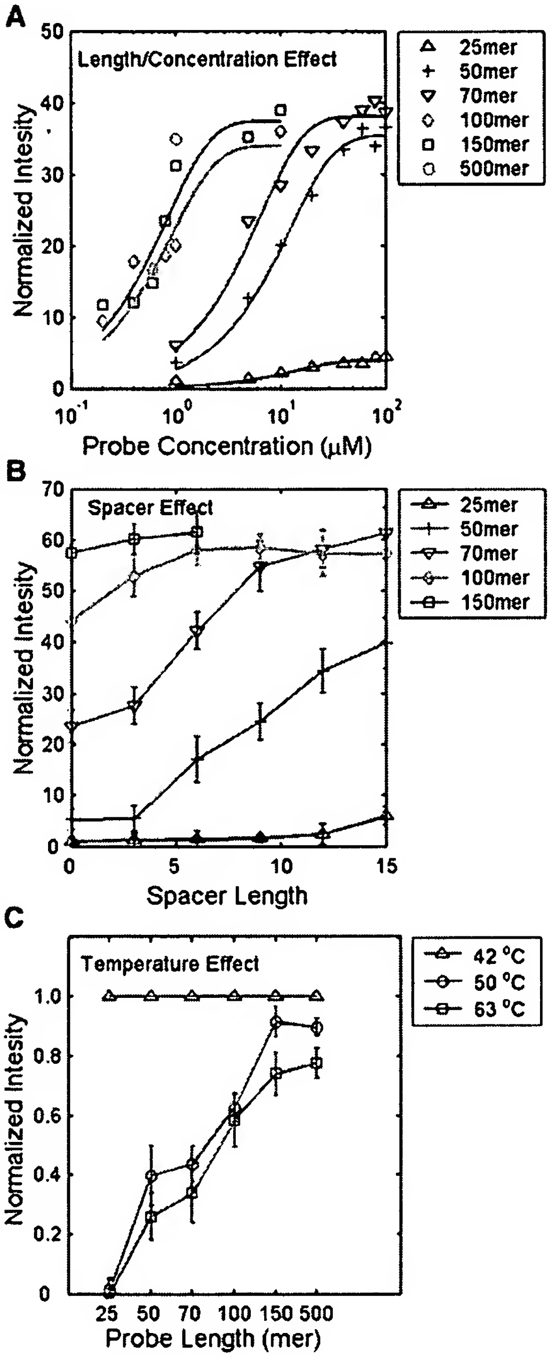


Figure 3. Assessment of discrepancies using different lengths of probes for hybridization. (A) Hybridization intensities at six probe lengths as a function of spotted probe concentrations. All the solid lines were fitted by regression analysis of the data points. Intensities were normalized to the intensity observed using the 25mer at a concentration of 1 μM. (B) Effect of spacer length on the mean hybridization intensity at five probe lengths. The basic single spacer unit had a length equivalent to a $(dT)_3$ oligonucleotide, so the scale is based on the equivalent number of nucleotides. Intensities were normalized to the intensity obtained using the 25mer probe with no added spacer. (C) Mean hybridization intensities at three hybridization/washing temperatures as a function of probe length. Intensities measured at a given probe length were normalized to that at 42°C. The above experiments were performed three times for signal averaging.

Temperature effect. The above experimental results indicated that 150mer or longer probes did not benefit from spacer addition, were long enough to overcome steric and diffusion limitation and could be treated like long cDNA probes in the hybridization reaction. This was verified by the experimental results shown in Figure 3C, in which hybrid stability was analyzed by the hybridization and post-hybridization washing temperature. We observed a reduction in hybridization intensity at higher hybridization/washing temperature for all of the six different probe lengths, but the effect started to reach a plateau at longer probe lengths. The data showed that 150 and 500mer probes gave similar results.

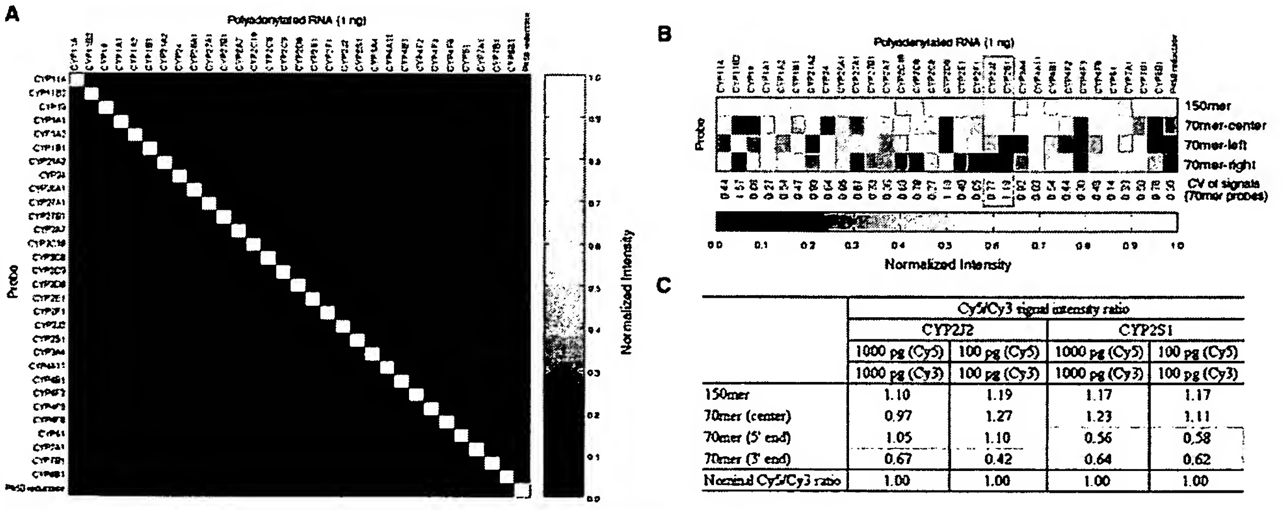
Array performance comparison of 70 and 150mer probes. Long oligonucleotide probes are becoming widely employed in commercial arrays for gene expression profiling (6,7,29). We therefore performed a series of experiments using the cytochrome P450 (CYP450) gene family to compare the array performance of 70 and 150mer probes. This gene family consists of many members with >80% sequence homology. The same criteria for avoiding cross-hybridization mentioned above were used to design gene-specific probes for the members of the gene family. Figure 4A shows the compiled results of 31 individual hybridization reactions. Each column shows the results of one Cy3-labeled, *in vitro*-transcribed CYP450 polyadenylated RNAs hybridized to an array containing the 31 150mer CYP450 gene probes spotted at 1 µM. No significant cross-hybridization was detected. These data demonstrate that gene-specific 150mer probes able to identify unique regions of a gene can be obtained using the computation approach.

In Figure 4B, the same set of 31 *in vitro*-transcribed CYP450 cDNAs was individually hybridized to arrays containing probes for the 31 CYP450 genes with 4 probes per gene. The 4 probes were the 150mer probe and 3 70mer probes, selected by the aforementioned probe design algorithm from the center, 5′- or 3′ end of the 150mer gene probe. All four probes had a GC content between 45 and 55% and a similar melting temperature. Similar to that described in Figure 3A, the 70 and 150mer probes were spotted on the P450 array at the optimal concentrations of 40 and 1 µM, respectively. In general, the 70mer CYP450 probes gave a poorer signal intensity than the corresponding 150mer probes. The average hybridization signal variation (CV of 0.6) calculated from the 31 sets of 3 70mer probes was the same as that in Figure 1B. In total, ~23% (21/93) of the 70mer CYP450 probes gave a hybridization signal <20% of the strongest signal for that gene.

The current microarray experimental protocol employs two-color fluorescence detection for differential expression ratio measurement. It has been reported that, using the same set of probes, the variation in the differential expression ratio is lower than that in the absolute hybridization signal (7), suggesting that multiple probes per gene may not be needed in most applications using differential expression ratio measurement. Our experimental results showed that this holds provided that an experimental pre-screening process is employed to select the best single probe with a good signal intensity for a gene. For instance, Figure 4C shows that the probes giving a good signal intensity for CYP2J2 and CYP2S1 (framed in Figure 4B), but not those giving a poor signal intensity, can provide accurate differential expression ratios.

DISCUSSION

In addition to sequence composition and probe length, target length is an important parameter in hybridization studies. Because of the poor hybridization efficiency of short probes, a protocol for RNA amplification using *in vitro* transcription (IVT) to generate labeled cRNA (30) is used to increase the



	Cy5/Cy3 signal intensity ratio			
	CYP2J2		CYP2S1	
	1000 pg (Cy5)	100 pg (Cy5)	1000 pg (Cy5)	100 pg (Cy5)
	1000 pg (Cy3)	100 pg (Cy3)	1000 pg (Cy3)	100 pg (Cy3)
150mer	1.10	1.19	1.17	1.17
70mer (center)	0.97	1.27	1.23	1.11
70mer (5′ end)	1.05	1.10	0.56	0.58
70mer (3′ end)	0.67	0.42	0.64	0.62
Nominal Cy5/Cy3 ratio	1.00	1.00	1.00	1.00

Figure 4. Array performance comparison of 70 and 150mer probes. (A) High specificity was seen using 31 CYP450 gene-specific 150mer probes. Each column represents the signal intensities of a CYP450 gene member hybridized to each of the 31 CYP450 probes. All signal intensities in each column were normalized to the brightest. Only the correct probe/target pairs yielded detectable signals. (B) Hybridization intensity variation using three 70mer CYP450 probes selected from within the corresponding 150mer gene-specific probe. (C) Probes with low hybridization sensitivity do not accurately measure the differential expression ratio. Lower signal intensity probes deviate greater from the nominal differential expression ratio of 1. All experiments were performed in triplicate for signal averaging. The Cy5/Cy3 intensity ratio was normalized by setting the arithmetic mean of the ratios of every spot on the array to 1.

amount of target molecules for hybridization. The process is followed by a fragmentation step to break down possible secondary structures in labeled cRNA target molecules and to increase the diffusion rate of the target molecules for microarray hybridization reactions with 20–30mer probes [(2,11), Amersham 30mer Uniset Bioarray protocol manual). For microarray experiments using long oligonucleotide probes, IVT is not used routinely and is only necessary when the amount of RNA is very low, e.g. RNA derived from tissue specimens.

The present study used a single hybridization protocol and probes of different lengths to examine their effects on gene expression measurement. To minimize experimental bias introduced by target length, a reverse transcription and labeling protocol using oligo(dT) and random primers (31) was employed. The protocol generated cDNA fragments ranging in length from 150 ± 60 to ~1000 bases, as revealed by slab gel electrophoresis (data not shown). The short target length facilitates the interaction of labeled cDNA with the probes during hybridization.

Previous literature reports on the analysis of oligonucleotide probe length have focused mainly on specificity and sensitivity (11,26) and, as far as we are aware of, this is the first study to examine the effect of probe length on hybridization intensity variation tested with different probes for a given gene and on measurement bias in gene expression profiling. We believe that all of these factors contribute to the discordant expression results observed using different microarray platforms. Since oligonucleotide hybridization efficiency is highly sequence dependent and cDNA probes are too long to avoid non-specific cross-hybridization, these two microarray platforms may not give concordant results. For instance, Tan *et al.* (32) evaluated three commercial microarrays, Agilent cDNA microarray, Affymetrix GeneChip and Amersham CodeLink Uniset Bioarray and found substantial differences using these three platforms. On the other hand, Barczak *et al.* (29) reported a good correlation ($r = 0.8$–0.9) between relative gene expression levels measured using a long oligonucleotide array (a single 70mer probe per gene) and Affymetrix GeneChip. On the basis of the reasonable assumption that the experimental protocols for these commercial microarrays have been optimized for gene expression profiling and that the additional RNA amplification (IVT) step employed in the short oligonucleotide probe microarray experiments does not introduce bias, a plausible explanation for the contradictory findings is sequence-dependent hybridization variation and measurement bias.

Barczak *et al.* (29) also discovered that two different collections of 70mer probes sometimes yielded dramatic signal differences for the two probes for the same gene. Our experimental results, shown in Figure 4B and C, agree with this finding and strongly suggest that a large-scale hybridization screening for the single best oligonucleotide probe per gene or multiple probes per gene is needed for making microarrays that yield reliable gene expression measurements. Large-scale probe screening was used in a recent study by Lucito *et al.* (33), who tested ~700 000 unique 70mer probes to find the probes giving the strongest hybridization signals. Our present study also demonstrated that a single 70mer or longer oligonucleotide probe for a gene could be sufficient for accurate expression measurement if the probe is validated experimentally.

The present study takes into account three factors to evaluate the performance of DNA microarrays of different probe lengths, investigates sequence-dependent effects and extends the study of the effect of probe length to beyond 500 bases. The experimental results suggest that gene-specific 150mer probes selected by a probe design computer program can minimize signal intensity variation and measurement bias to such an extent that prior experimental screening is not necessary, i.e. the hybridization properties of 150mer probes are similar to those of long cDNA probes and 150mer probes can be used for gene-specific expression measurements.

In summary, nucleic acid hybridization involves a multitude of variables not yet predictable by computation alone. A good probe design strategy is essential for accurate expression measurement. This study focused on characterizing the effect of probe length and the number of probes per gene on accurate microarray measurement of gene expression. Our data suggest that probes ~150mer in length are optimal for this purpose. Short or long oligonucleotide probes can also work well if the probes are validated by experimental hybridization selection or if multiple probes per gene are used. We believe that concordant results can be obtained using different microarray platforms if all the negative effects are minimized.

ACKNOWLEDGEMENTS

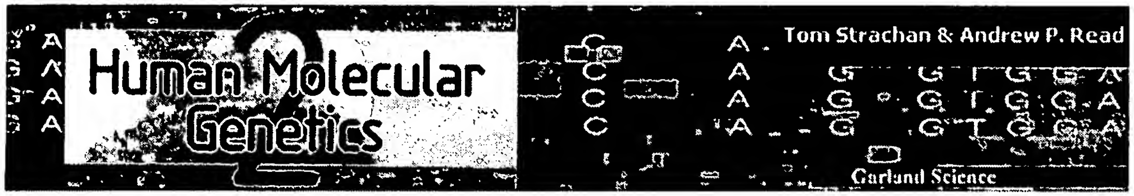
We thank Hung-Hui Chen, Ping-Ann Chen, Li-Jen Huang and Bi-Yun Lee for preparing the oligonucleotides and the gene-specific probes. We also thank Dr Tom Barkas for reviewing the manuscript for us. This project was supported by grants from the National Research Program for Genome Medicine (NSC 92-3112-B-001-013; NSC 93-3112-B-001-013-Y) of the National Science Council and a fellowship to C-.C.C. from the Medical Biotechnology Program of the Academia Sinica.

REFERENCES

1. Schena,M., Shalon,D., Davis,R.W. and Brown,P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.
2. Lockhart,D.J., Dong,H., Byrne,M.C., Follettie,M.T., Gallo,M.V., Chee,M.S., Mittmann,M., Wang,C., Kobayashi,M., Horton,H. *et al.* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.
3. Xu,W., Bak,S., Decker,A., Paquette,S.M., Feyereisen,R. and Galbraith,D.W. (2001) Microarray-based analysis of gene expression in very large gene families: the cytochrome P450 gene superfamily of *Arabidopsis thaliana*. *Gene*, **272**, 61–74.
4. Evertsz,E.M., Au-Young,J., Ruvolo,M.V., Lim,A.C. and Reynolds,M.A. (2001) Hybridization cross-reactivity within homologous gene families on glass cDNA microarrays. *Biotechniques*, **31**, 1182, 1184, 1186.
5. Modrek,B. and Lee,C. (2002) A genomic view of alternative splicing. *Nature Genet.*, **30**, 13–19.
6. Kane,M.D., Jatkoe,T.A., Stumpf,C.R., Lu,J., Thomas,J.D. and Madore,S.J. (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res.*, **28**, 4552–4557.
7. Hughes,T.R., Mao,M., Jones,A.R., Burchard,J., Marton,M.J., Shannon,K.W., Lefkowitz,S.M., Ziman,M., Schelter,J.M., Meyer,M.R. *et al.* (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.*, **19**, 342–347.
8. Wright,M.A. and Church,G.M. (2002) An open-source oligomicroarray standard for human and mouse. *Nat. Biotechnol.*, **20**, 1082–1083.

9. Holloway,A.J., Van Laar,R.K., Tothill,R.W. and Bowtell,D.D. (2002) Options available-from start to finish-for obtaining data from DNA microarrays II. *Nature Genet.*, **32** (Suppl 2), 481–489.
10. Kuo,W.P., Jenssen,T.K., Butte,A.J., Ohno-Machado,L. and Kohane,I.S. (2002) Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics*, **18**, 405–412.
11. Relogio,A., Schwager,C., Richter,A., Ansorge,W. and Valcarcel,J. (2002) Optimization of oligonucleotide-based DNA microarrays. *Nucleic Acids Res.*, **30**, e51.
12. Tijssen,P. (1993) *Hybridization With Nucleic Acid Probes, Part I: Theory and Nucleic Acid Preparation.* Elsevier Science Publisher, NY.
13. Selinger,D.W., Cheung,K.J., Mei,R., Johansson,E.M., Richmond,C.S., Blattner,F.R., Lockhart,D.J. and Church,G.M. (2000) RNA expression analysis using a 30 base pair resolution *Escherichia coli* genome array. *Nat. Biotechnol.*, **18**, 1262–1268.
14. Li,F. and Stormo,G.D. (2001) Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics*, **17**, 1067–1076.
15. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
16. Walter,A.E., Turner,D.H., Kim,J., Lyttle,M.H., Muller,P., Mathews,D.H. and Zuker,M. (1994) Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc. Natl Acad. Sci. USA*, **91**, 9218–9222.
17. Cheng,J.Y., Chen,H.H., Kao,Y.S., Kao,W.C. and Peck,K. (2002) High throughput parallel synthesis of oligonucleotides with 1536 channel synthesizer. *Nucleic Acids Res.*, **30**, e93.
18. Yue,H., Eastman,P.S., Wang,B.B., Minor,J., Doctolero,M.H., Nuttall,R.L., Stack,R., Becker,J.W., Montgomery,J.R., Vainer,M. *et al.* (2001) An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression. *Nucleic Acids Res.*, **29**, e41.
19. Taylor,J.R. (1982) *An Introduction to Error Analysis: The Statistical Study of Uncertainties in Physical Measurement.* University Science Books, Mill Valley, CA.
20. Berger,J.O. (1985) *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, NY.
21. DeGroot,M.H. (1970) *Optimal Statistical Decisions.* McGraw-Hill, NY.
22. Smith,J.Q. (1988) *Decision Analysis.* Chapman and Hall, London, UK.
23. Spiring,F.A. (1993) The reflected normal loss function. *Can. J. Stat.*, **21**, 321–330.
24. Hegde,P., Qi,R., Abernathy,K., Gay,C., Dharap,S., Gaspard,R., Hughes,J.E., Snesrud,E., Lee,N. and Quackenbush,J. (2000) A concise guide to cDNA microarray analysis. *Biotechniques*, **29**, 548–550, 552–554, 556.
25. Guo,Z., Liu,Q. and Smith,L.M. (1997) Enhanced discrimination of single nucleotide polymorphisms by artificial mismatch hybridization. *Nat. Biotechnol.*, **15**, 331–335.
26. Guo,Z., Guilfoyle,R.A., Thiel,A.J., Wang,R. and Smith,L.M. (1994) Direct fluorescence analysis of genetic polymorphisms by hybridization with oligonucleotide arrays on glass supports. *Nucleic Acids Res.*, **22**, 5456–5465.
27. Shchepinov,M.S., Case-Green,S.C. and Southern,E.M. (1997) Steric factors influencing hybridisation of nucleic acids to oligonucleotide arrays. *Nucleic Acids Res.*, **25**, 1155–1161.
28. Southern,E., Mir,K. and Shchepinov,M. (1999) Molecular interactions on microarrays. *Nature Genet.*, **21**, 5–9.
29. Barczak,A., Rodriguez,M.W., Hanspers,K., Koth,L.L., Tai,Y.C., Bolstad,B.M., Speed,T.P. and Erle,D.J. (2003) Spotted long oligonucleotide arrays for human gene expression analysis. *Genome Res.*, **13**, 1775–1785.
30. Shannon,K.W. (2000) Method for linear mRNA amplification. US patent no. 6,132,997.
31. Bowtell,D. and Sambrook,J. (2002) *DNA Microarrays: A Molecular Cloning Manual.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
32. Tan,P.K., Downey,T.J., Spitznagel,E.L.,Jr, Xu,P., Fu,D., Dimitrov,D.S., Lempicki,R.A., Raaka,B.M. and Cam,M.C. (2003) Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.*, **31**, 5676–5684.
33. Lucito,R., Healy,J., Alexander,J., Reiner,A., Esposito,D., Chi,M., Rodgers,L., Brady,A., Sebat,J., Troge,J. *et al.* (2003) Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Res.*, **13**, 2291–2305.

Exhibit 3



Short Contents | Full Contents Other books @ NCBI

Navigation

About this book

6. PCR, DNA sequencing and *in vitro* mutagenesis

➔ 6.1. Basic features of PCR

6.2. Applications of PCR

6.3. DNA sequencing

6.4. *In vitro* site-specific mutagenesis

Further reading

Figures

Figure 6.1. PCR is an in vitro...

Figure 6.2. PCR primer design.

Figure 6.3. Cloning of PCR products in...

Boxes

Box 6.1. Proofreading by DNA polymerase-associated 3′...

Search

Go

● This book ○ All books

○ PubMed

Human Molecular Genetics 2 ➔ 6. PCR, DNA sequencing and *in vitro* mutagenesis

6.1. Basic features of PCR

The polymerase chain reaction (PCR) has revolutionized molecular genetics by permitting rapid cloning and analysis of DNA. Since the first reports describing this new technology in the mid 1980s, there have been numerous applications in both basic and clinical research. Two other fundamental technologies are DNA sequencing and *in vitro* mutagenesis, both of which can be accomplished using PCR-based and non PCR-based methods.

6.1.1. PCR is a cell-free method of DNA cloning

The standard PCR reaction: selective DNA amplification

PCR is a rapid and versatile *in vitro* method for amplifying defined target DNA sequences present within a source of DNA. Usually, the method is designed to permit *selective amplification* of a specific target DNA sequence(s) within a heterogeneous collection of DNA sequences (e.g. total genomic DNA or a complex cDNA population). To permit such selective amplification, some prior DNA sequence information from the target sequences is required. This information is used to design two oligonucleotide primers (amplimers) which are specific for the target sequence and which are often about 15–25 nucleotides long. After the primers are added to denatured template DNA, they bind specifically to complementary DNA sequences at the target site. In the presence of a suitably heat-stable DNA polymerase and DNA precursors (the four deoxynucleoside triphosphates, dATP, dCTP, dGTP and dTTP), they initiate the synthesis of new DNA strands which are complementary to the individual DNA strands of the target DNA segment, and which will overlap each other (*Figure 6.1*).

The PCR is a chain reaction because newly synthesized DNA strands will act as templates for further DNA synthesis in subsequent cycles. After about 25 cycles of DNA synthesis, the products of the PCR will include, in addition to the starting DNA, about 10^5 copies of the specific target sequence, an amount which is easily visualized as a discrete band of a specific size when submitted to agarose gel electrophoresis. A heat-stable DNA polymerase is used because the reaction involves sequential cycles composed of three steps:

i. Denaturation, typically at about 93–95°C for human genomic DNA.
ii. Reannealing at temperatures usually from about 50°C to 70°C depending on the T_m (see Section 5.2.1) of the

expected duplex (the annealing temperature is typically about 5°C below the calculated T_m).

iii. DNA synthesis, typically at about 70–75°C.

Suitably heat-stable DNA polymerases have been obtained from microorganisms whose natural habitat is hot springs. For example, the widely used *Taq* DNA polymerase is obtained from *Thermus aquaticus* and is thermostable up to 94°C, with an optimum working temperature of 80°C. ↑ TOP

Specificity of amplification and primer design

The specificity of amplification depends on the extent to which the primers can recognize and bind to sequences other than the intended target DNA sequences. For complex DNA sources, such as total genomic DNA from a mammalian cell, it is often sufficient to design two primers about 20 nucleotides long. This is because the chance of an accidental perfect match elsewhere in the genome for either one of the primers is extremely low, and for both sequences to occur by chance in close proximity in the specified direction is normally exceedingly low. Although conditions are usually chosen to ensure that only strongly matched primer-target duplexes are stable, spurious amplification products can nevertheless be observed. This can happen if one or both chosen primer sequences contain part of a repetitive DNA sequence, and primers are usually designed to avoid matching to known repetitive DNA sequences, including large runs of a single nucleotide (*Figure 6.2*).

Accidental matching at the $3'$ end of the primer is critically important: spurious products may derive from substantially mismatched primer-target duplexes unless the $3'$ end of the primer shows perfect matching. Several strategies can be adopted to optimize reaction specificity:

- **Nested primers**. The products of an initial amplification reaction are diluted and used as the target DNA source for a second reaction in which a different set of primers is used, corresponding to sequences located close, but internal, to those used in the first reaction.
- **Hot-start PCR**. Mixing of all PCR reagents prior to an initial heat denaturation step allows more opportunity for nonspecific binding of primer sequences. To reduce this possibility, one or more components of the PCR are physically separated until the first denaturation step. A popular approach is to use a specially formulated wax bead designed to fit snugly within a PCR reaction tube. The reaction components minus the enzyme and reaction buffer are added to the tube followed by the molten wax bead which floats on top and then solidifies on cooling. The thermostable polymerase is then added with buffer. At the initial denaturation step the wax melts again and rises to the surface causing all the reaction components to come into contact with each other.
- **Touch-down PCR**. Most thermal cyclers can be programed to perform runs in which the annealing temperature is lowered incrementally during the PCR cycling from an initial value above the expected T_m to a value below the T_m. By keeping the stringency of hybridization initially very high, the formation of spurious products is discouraged, allowing the expected sequence to predominate.

↑ TOP

DNA labeling by PCR

The standard PCR reaction can be modified to permit incorporation of labeled nucleotides. Two methods are commonly used:

- **Standard PCR-based DNA labeling**. The PCR reaction is modified to include one or more labeled nucleotide precursors which become incorporated into the PCR product throughout its length.
- **Primer-mediated $5'$ end labeling**. PCR is conducted using a primer in which a labeled group is attached to the $5'$ end. As PCR proceeds the primer with its $5'$ end-label is incorporated into the PCR product. This method is often used with fluorophore labels during DNA sequencing (see legend to *Figure 6.18*) and is an example of a general PCR mutagenesis method known as **$5'$ add-on mutagenesis** which has many applications (see Section 6.4.2 and *Figure 6.20A*).

↑ TOP

6.1.2. The major advantages of PCR as a cloning method are its rapidity, sensitivity and robustness

Because of its simplicity, PCR is a popular technique with a wide range of applications which depend on essentially three major advantages of the method.

Speed and ease of use

DNA cloning by PCR can be performed in a few hours, using relatively unsophisticated equipment. Typically, a PCR reaction consists of 30 cycles containing a denaturation, synthesis and reannealing step, with an individual cycle typically taking 3–5 min in an automated thermal cycler. This compares favorably with the time required for cell-based DNA cloning, which may take weeks. Clearly, some time is also required for designing and synthesizing oligonucleotide primers, but this has been simplified by the availability of computer software for primer design and rapid commercial synthesis of custom oligonucleotides. Once the conditions for a reaction have been tested, the reaction can then be repeated simply. ↑ TOP

Sensitivity

PCR is capable of amplifying sequences from minute amounts of target DNA, even the DNA from a single cell (Li et al., 1988). Such exquisite sensitivity has afforded new methods of studying molecular pathogenesis and has found numerous applications in forensic science, in diagnosis, in genetic linkage analysis using single-sperm typing and in molecular paleontology studies, where samples may contain minute numbers of cells. However, the extreme sensitivity of the method means that great care has to be taken to avoid contamination of the sample under investigation by external DNA, such as from minute amounts of cells from the operator. ↑ TOP

Robustness

PCR can permit amplification of specific sequences from material in which the DNA is badly degraded or embedded in a medium from which conventional DNA isolation is problematic. As a result, it is again very suitable for molecular anthropology and paleontology studies, for example the analysis of DNA recovered from archaeological remains. It has also been used successfully to amplify DNA from formalin-fixed tissue samples, which has important applications in molecular pathology and, in some cases, genetic linkage studies. ↑ TOP

6.1.3. The major disadvantages of PCR are the general requirement for prior target sequence information, short size and limiting amounts of product, and infidelity of DNA replication

Despite its huge popularity, PCR has certain limitations as a method for selectively cloning specific DNA sequences.

Need for target DNA sequence information

In order to construct specific oligonucleotide primers that permit selective amplification of a particular DNA sequence, some prior sequence information is necessary. This normally means that the DNA region of interest has been partly characterized previously, often following cell-based DNA cloning. However, a variety of techniques have been developed that reduce or even exclude the need for prior DNA sequence information concerning the target DNA, when certain aims are to be met. For example, previously uncharacterized DNA sequences can sometimes be cloned using PCR with degenerate oligonucleotides if they are members of a gene or repetitive DNA family at least one of whose members has previously been characterized. In some cases, PCR can be used effectively without any prior sequence information concerning the target DNA to permit *indiscriminate amplification* of DNA sequences from a source of DNA that is present in extemely limited quantities (Section 6.2.4). Therefore, although PCR can be applied to ensure whole genome amplification, it does not have the advantage of cell-based DNA cloning in offering a way of separating the individual DNA clones comprising a genomic DNA library. ↑ TOP

Short size and limiting amounts of PCR product

A clear disadvantage of PCR as a DNA cloning method has been the size range of the DNA sequences that can be cloned. Unlike cell-based DNA cloning where the size of cloned DNA sequences can approach 2 Mb (Section 4.3.4), reported DNA sequences cloned by PCR have typically been in the 0.1–5 kb size range, often at the lower end of this scale. Although small segments of DNA can usually be amplified easily by PCR, it becomes increasingly more difficult to obtain efficient amplification as the desired product length increases. Recently, however, conditions have been identified for effective amplification of longer targets, including a 42-kb product from the bacteriophage λ genome. Often, the conditions for long range PCR involve a combination of modifications to standard conditions with a two-polymerase system. This provides optimal levels of DNA polymerase and $3' \rightarrow 5'$ exonuclease activity which serves as a proofreading mechanism (see *Box 6.1*).

The amount of PCR product obtained in a single reaction is also much more limited than the amount that can be obtained using cell-based cloning where scale-up of the volumes of cell cultures is possible. The efficiency of a PCR reaction will vary from template to template and according to various factors that are required to optimize the reaction but typically only comparatively small amounts of product are achieved. ↑ TOP

Infidelity of DNA replication

Cell-based DNA cloning involves DNA replication *in vivo,* which is associated with a very high fidelity of copying because of proofreading mechanisms (see *Box 6.1*). However, when DNA is replicated *in vitro* the copying error rate is considerably greater. Of the heat-stable DNA polymerases required for PCR, the most widely used is *Taq* DNA polymerase derived from *T. aquaticus*. This DNA polymerase, however, has no associated $3' \rightarrow 5'$ exonuclease to confer a proofreading function, and the error rate due to base misincorporation during DNA replication is rather high: for a 1 kb sequence that has undergone 20 effective cycles of duplication, approximately 40% of the new DNA strands synthesized by PCR using this enzyme will contain an incorrect nucleotide resulting from a copying error. This means that, even if the PCR reaction involves amplification of a single DNA sequence, the final product will be a mixture of extremely similar, but not identical DNA sequences.

Despite the errors due to replication *in vitro*, DNA sequencing of the total PCR product may give the correct sequence. This is because, although individual DNA strands in the PCR product often contain incorrect bases, the incorporation of incorrect bases is essentially random. As a result, *for each base position,* the contribution of one incorrect base on one or more strands is overwhelmed by the contributions from the huge majority of strands which will have the correct sequence. What it does mean, however, is that further analysis of the product may be difficult. If the PCR product is to be cloned in cells (e.g. to facilitate DNA sequencing or to permit functional studies in a cell-based expression system), transformation selects for a single molecule, and the cell clones chosen to be amplified will contain identical molecules, each the same as a single starting molecule which may well have the incorrect DNA sequence because of a copying error during PCR amplification. As a result, several individual clones may need to be sequenced in order to determine the correct (consensus) sequence, before selecting one with the authentic sequence for subsequent experiments.

More recently, the problem of infidelity of DNA replication during the PCR reaction has been considerably reduced by using alternative heat-stable DNA polymerases which have associated $3'$–$5'$ exonuclease activity. For example, the *Pyrococcus furiosus* DNA polymerase is becoming more widely used because of the proofreading conferred by its associated $3'$–$5'$ exonuclease activity (Cline *et al.*, 1996). The resulting PCR product has a much lower level of mutations introduced by copying errors: for a 1 kb segment of DNA that has undergone 20 effective cycles of duplication, about 3.5% of the DNA strands in the product carry an altered base. ↑ TOP

6.1.4. Cell-based cloning of PCR amplification products is often required to permit subsequent structural and functional studies

The amount of material that can be cloned in a single PCR reaction is limited, and it is time-consuming and expensive to repeat the same PCR reaction many times to achieve large quantities of the desired DNA. In addition, the PCR product may not be in a suitable form that will permit some subsequent studies. As a result, it is often convenient to clone the PCR product in a cell-based cloning system in order to obtain large quantities of the desired DNA and to permit a variety of analyses. As described in the previous section, it is important to verify that the sequence of the cloned product is representative of the original PCR product.

Various plasmid cloning systems are used to propagate PCR-cloned DNA in bacterial cells. Once cloned, the insert can be cut out using suitable restriction nucleases and transferred into other plasmids which may have specialized usages in permitting expression to give an RNA product, or to provide large quantities of a protein, etc. Several thermostable polymerases including *Taq* DNA polymerase have a terminal deoxynucleotidyl transferase activity which selectively modifies PCR-generated fragments by adding a single nucleotide, generally adenine, to the 3′ ends of amplified DNA fragments. The resulting overhangs can make it difficult to clone PCR products and a variety of approaches are commonly used to facilitate cloning, including the use of vectors with overhanging T residues in their cloning site polylinker and the use of 'polishing' enzymes such as T4 polymerase or *Pfu* polymerase which can remove the overhanging single nucleotides (*Figure 6.3*).

↑ TOP

Exhibit 4

GENOMES 2 SECOND EDITION

Garland Science

T.A. BROWN

Short Contents | Full Contents

Other books @ NCBI

Navigation

About this book

2. Studying Genomes

5. Mapping Genomes

5.1. Genetic and Physical Maps

5.2. Genetic Mapping

5.3. Physical Mapping

Study Aids For Chapter 5

References

Further Reading

Search

Go

◉ This book ○ All books

○ PubMed

Genomes 2 → 2. Studying Genomes → 5. Mapping Genomes → 5.3. Physical Mapping

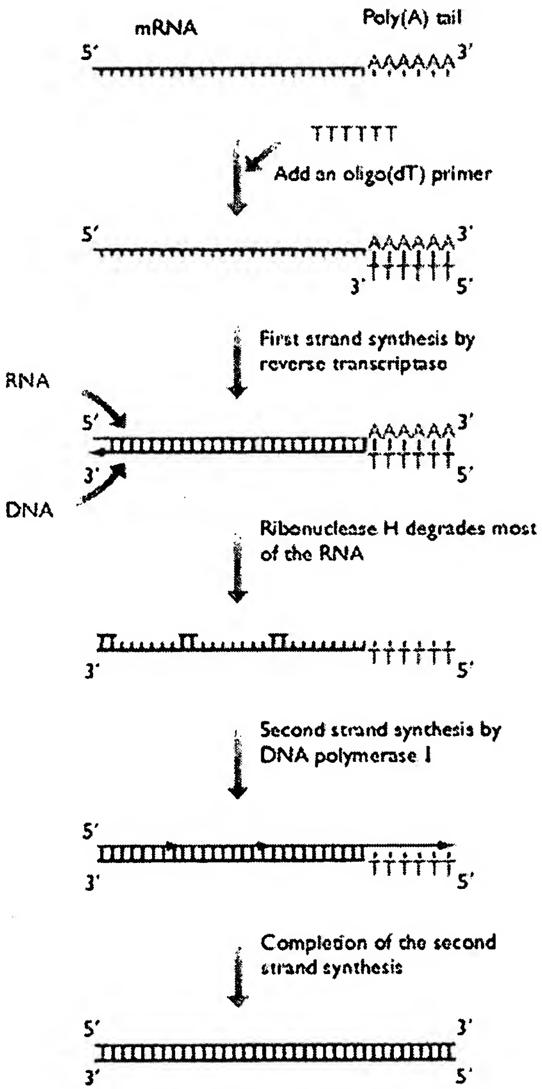

mRNA
Poly(A) tail
5'
AAAAAA
3'
TTTTTT
Add an oligo(dT) primer
First strand synthesis by reverse transcriptase
RNA
DNA
Ribonuclease H degrades most of the RNA
Second strand synthesis by DNA polymerase I
Completion of the second strand synthesis

Figure 5.32. One method for preparing cDNA. Most eukaryotic mRNAs have a poly(A) tail at their 3′ end

(Section 10.1.2). This series of A nucleotides is used as the priming site for the first stage of cDNA synthesis, carried out by reverse transcriptase - a DNA polymerase that copies an RNA template (Section 4.1.1). The primer is a short synthetic DNA oligonucleotide, typically 20 nucleotides in length, made up entirely of Ts (an 'oligo(dT)' primer). When the first strand synthesis has been completed, the preparation is treated with ribonuclease H, which specifically degrades the RNA component of an RNA-DNA hybrid. Under the conditions used, the enzyme does not degrade all of the RNA, instead leaving short segments that prime the second DNA strand synthesis reaction, this one catalyzed by DNA polymerase I.